DE-centralised Cloud labs fOr inDustrialisation of Energy materials

DELIVERABLE 5.2

Data Management Plan

Organisation: Forschnugszentrum Juelich GmbH

Main authors: Kourosh Malek, Michael Eikerling

Date: 30.06.2024

DELIVERABLE 5.2 – VERSION n°1

WORK PACKAGE 5

Funded by the European Union

Nature of the deliverable		
R	Document, report	X
DEM	Demonstrator, pilot, prototype	
DMP	Data Management Plan	
OTHER	Software, technical diagram, etc	

Dissemination level		
PU	Public (*fully open*)	
SEN	Sensitive (*limited under the conditions of the Grant Agreement*)	X
EU CI	EU Classified (*eu-restricted, eu-confidential, eu-secret under Decision 2015/444*)	

Deliverable No.	5.2
Dissemination level	Public
Work Package	WP5: Project, Innovation and IPR Management
Task	T5.3: Data Management
Lead beneficiary	FZJ
Contributing beneficiaries	ALL
Due date of deliverable	M6 – 31/05/2024
Actual submission date	15.06.2024

Quality procedure			
Date	Version	(potential) Reviewers	Comments
29.07.2024	V1	Michael Eikerling	
23.04.2024	V2		
	V3		
17.05.2024	V3		
23.05.2024	V4		
31.05.2024	V4		

Acknowledgements
This report is part of the 57 deliverables from the project "DECODE" which has received funding from the European Union's Horizon Europe research and innovation program under grant agreement N° 101135537. More information on the project can be found at www.decode-energy.eu.

Document objective and executive summary

At the outset of the DECODE project, this data management plan (DMP) is developed to ensure project data and translated knowledge thereof, are complied with FAIR principles. In order to provide highly detailed rules for managing data, this plan will be regularly reviewed and adjusted as required every six months throughout the project's duration. Data collected in DECODE, either newly generated or existing data, will be used for Artificial Intelligence approaches to develop new AI/ML models or fine-tune existing models. As it will mainly consist in analysing and gathering data, it does not raise concerns regarding human rights, or personal aspects, and will not have any negative impacts on DECODE commitment to high standard in ethics and data privacy.

The present deliverable (D5.2) should be read in line with the D5.1 on the "Ethics Requirements". In addition to that, the policy requirements and obligations of DECODE in view of collected scientific, technical, public, pre-commercial, commercial or personal data are addressed in the framework of various Work Packages as defined in the original proposal, Part B. A first version of the plan is provided here in this report:

DECODE collects experimental and technical data related to research and development and demonstration, excluding personal data. Examples of **planned collected data** include materials selection and their intrinsic properties, fabrication and testing conditions, data on imaging, characterisation, modelling and diagnosis of components and cells, performance and lifetime, final device characterisations, materials and energy balances for the LCA inventory, website, and publication data. Partners will use this data to monitor KPIs and track progress made in the project.

In order to ensure the **findability of data/research outputs**, each partner's repositories will be used to store, and backup data collected/generated during the DECODE project, along with corresponding metadata for retrieval purposes. Partners that collect/generate data in DECODE have internal procedures in place, in accordance with national and EU laws (including TECH). Relevant data will also be stored on a secure project SharePoint to be shared with the consortium. Data that is not submitted to IPR and DECODE outcomes will mainly be exploited in publications with their own digital object identifier, in line with open science practices. These data will also be available on the project website and social media, as well as stored on a trusted repository such as Zenodo or the *European Open Science Cloud*, in a standard format that complies with open research data requirements.

DECODE project partners will have full **access to collected/generated data** using the secure SharePoint. They also have access to all pre-trained AI models, physical models, modules and underlying training or validation/testing data using a unified cloud platform. Upon legitimate request, partners may grant specific access to third parties with respective partners' approval, in accordance with the IPR strategy detailed in the CA.

Data will be converted to **interoperable formats** whenever possible, such as those supported by the Microsoft package and PDF, to ensure that each partner can access and process it, and it can be handled in accordance with the applicable law in each state, including GDPR. The **FZJ** as the coordinator appoints a dedicated data protection officer for handling all GDPR matters. A naming convention will be established for each DECODE document and deliverable.

Data will remain usable after the end of DECODE to plan ongoing impact, especially by the **FZJ** as the coordinator and interested third parties. The consortium will have authorship of **data and documents for re-use**. Dissemination and communication material will be open to re-use by third parties as soon as they are made available free of charge on the project website and social media, for an unlimited period, together with data published in scientific articles.

Each partner involved in the collection and management of data will use their usual practices for **curation and preservation of data**.

DECODE ensures that (potentially) sensitive data, including technical as well as personal data, will be treated following ethical data handling practices, as stipulated in national and European data privacy regulations. Personal data are processed and protected in accordance with European data protection regulations (GDPR) as well as national regulations, as required by partner countries. The individuals whose data are being collected for project-related purposes, will be informed about the intended uses of their data and their written consent for these uses will be obtained. They will be informed about their rights as data subjects and about the steps to be performed to unsubscribe. The collected data will be stored on a project internal, secured and password-protected servers. The data will be deleted immediately after each event or at the completion of the project, depending on the type of consent given by the individuals whose data are concerned.

Forschungszentrum Juelich GmbH (FZJ), the coordinator**,** will be responsible for data management in DECODE, and will in particular, responsible to update this DMP on a regular basis. This version of DMP will detail the overall data management policy. In particular, this report is specifying which kind of data will be generated by the project, how it will be exploited or made accessible, how it will be secured, and how it will be curated and preserved. This DMP will be improved with all partners' feedback, producing a first version at M8, shortly after first consortium meeting, which will be further revised in M8 and every six months if needed beyond that point. The DMP will also be validated by non-EU partner. Also, **DECODE** has developed a database accessible to all partners (secure project SharePoint) at the beginning of the project, and which will be used all along DECODE. Finally, this DMP ensure that the description of the anonymisation or pseudonymisation techniques used will be kept on file.

Table of Contents

List of Figures

Abbreviations

CA: Consortium agreement
DMP: Data management plan
EMMO: European Materials Modelling Ontology
FAIR: Findable, Accessible, Interoperable, and Re-usable
FZJ: Forschungszentrum Juelich GmbH
GA: Grant Agreement
GDPR: European general data protection regulations
LCA: Life cycle assessment
LLM: Large Language models
POPD: Processing of Personal Data

Introduction

The DECODE project recognises the value of regulating research and technical data assets, as well as implementation of a robust personal data management across various activities. This document presents the first update of the initial Data Management Plan (DMP) of DECODE. This report is created particularly considering the data system put in place and presented in the project's first review meeting held on June 6 -7, 2024. It also includes the initial inputs received from the partners, particularly related to the data access and standardization across different partner labs. The report is comprising (scientific) data management, but also includes document management (e.g., reports, presentations, agreements) and knowledge management, Intellectual Property (IP), process flows, data security protocol, and managing AI-ready data.

Technical and scientific data management

R&D data assets are poorly utilized for developing new energy materials within emerging clean technologies. While there are open data, physical models, and data-driven tools, the basis of technical data is highly fragmented. The material integration process, beyond lab-scale materials synthesis and characterization, has been a time-consuming and expensive process, mainly due to lack of curated and systematic data. The discovery-to-device pipeline of new materials is thus a complex task to follow which is largely driven unstructured and complex data (type, category, sources) from various experimental, synthesis, modelling and simulation sources.

This report articulates a practical material-to-device data pipeline and its associated cloud infrastructure. The cloud-enabled infrastructure (ViMi Labs) facilitates access to AI-ready data sources and predictive models, provides lab-to-lab access for accelerating development, and early-prediction of materials and component in-cell performance.

Most of the AI/ML and energy storage analysis models that are discussed and implemented in DECODE have already been developed and previously published external to project DECODE. These models, however, have been re-built to adapt to the new model containers in course of DECODE project, ensuring their scalability and multi-user performance.

Personal Data in DECODE project

Personal data refers to any information, in any form, including document, spread sheets, tables, in binary, ascii, encrypted or in text formats, relating to an identified or identifiable natural person. This can include direct identifiers such as names and addresses, as well as indirect identifiers like IP addresses or pseudonyms that, when combined with other data, can identify an individual. Personal data can be processed and protected in accordance with European data protection regulations (GDPR) as well as national regulations, as required by partner countries.[1] The individuals whose data are being collected for project-related purposes, will be informed about the intended uses of their data and their written consent for these uses will be obtained. They will be informed about their rights as data subjects and about the steps to be performed to unsubscribe. The collected data will be stored on a

[1] European data protection regulations (GDPR), General Legal Text, https://gdpr-info.eu/

project internal, secured and password-protected servers. The data will be deleted immediately after each event or at the completion of the project, depending on the type of consent given by the individuals whose data are concerned.

Objectives of the Report

This DMP represents the project's most recent view on the datasets and exchanges in the project and between the project and the "outside world" and will be updated as the project progresses since not all data or potential uses are clear at this stage of the project. This DMP will be updated every time that significant changes arise, such as (but not limited to) new data types or data classes or sources are added; when certain changes in consortium policies occur (e.g., innovation potential, decision to file for a patent); changes in consortium composition and external partners as well at any new exploitation events. If no significant changes arise over the course of the project the DMP will be updated as a minimum in time with the periodic report in month 18, 36 and the final report in month 48.

This DMP is a tool for identifying and managing all the data in the project, to identify its origin, its content, the responsible person (data manager), and the servers where they will be hosted. In this respect, the document contains an overview of datasets collected/generated within the project indicating the Work Packages (WPs) and tasks (Ts) dealing with them. The DMP also serves as a guideline to make the project data Findable, Accessible, Interoperable, and Re-usable (FAIR) according to the Horizon Europe Data Management (DM) principles, which will be both useful for internal purposes in the project consortium and for the external community (regarding the publicly available part of the project data). The DMP provides first elements on how the FAIR-ness approach will be realised in practice. More details will be provided in subsequent versions as the work proceeds in the WPs.

In alignment with the "Guidelines on FAIR Data Management in Horizon Europe[2]" the present DMP describes the data management life cycle for the data to be collected, processed and/or generated by the project. As part of making research data FAIR, the present DMP includes information on the types of research data that will be generated or collected during the project, the standards that will be used, how the research data will be preserved and what parts of the datasets will be shared for verification or re-used.

The DECODE project will also collect personal data. In this context, we will therefore strictly consider the regulations of the participating countries on data privacy-related issues and the EU Directive on the Protection of Personal Data, as well as the General Data Protection Regulation (GDPR). More details regarding Processing of Personal Data (POPD) are given in D5.1. The DECODE DMP has been prepared with the aid of the guidelines provided in the EU online manual.[3]

Accordingly, to manage the data generated within the project or collected for the purposes of this project, and in line with the rules laid down in the Grant Agreement (GA) and DOA-part B, the project intends to use ViMi Labs platform[4] and its extended application/models marketplace to adopt a de-centralized approach. This approach will enable any user to access, mine, exploit, reproduce and disseminate data free of charge. It will also provide information about tools and instruments necessary

[2] European Commission, Directorate-General for Research & Innovation, H2020 Programme, Guidelines on FAIR Data Management in Horizon 2020, Version 3.0, 26 July 2016.

[3] https://webgate.ec.europa.eu/funding-tenders-opportunities/display/OM/Online+Manual

[4] Vimilabs.com

for validating the results (providing the tools and instruments themselves whenever possible, or providing at least information, via the chosen repository, about the tools necessary for validating the results, such as specialised software or software code, algorithms, analysis protocols, and similar.

Gap analysis of research data management in DECODE

In this section, we address the gaps in availability of practical material-to-device data infrastructure within energy materials domain. Characterization and diagnosis are a key capability to support deployment of energy materials [3,4]. For industrialization, having well defined technical specification of energy materials is critical because it enables standardization of input materials (catalyst, battery electrodes, electrocatalyst support, ionomer and membrane, solid electrolytes). Significant gaps remain in the fuel cells, water electrolysis and other electrochemical cells. Quality assurance and control during manufacturing cell and stack relies on suitable techniques to either pre-qualify materials or sampling during the manufacturing process or perform testing at the stack level. These samplings/QA are required by the target applications (e.g., 6 sigma certifications to be met by the automotive industry). New approaches are discussed in big data management, energy materials data space, artificial intelligence (AI), cloud software infrastructure, and de-centralizing cloud labs for modeling-characterization data pipelines.

Limitations in data infrastructure and data models

The infrastructure for intelligent data management and analytics is generally underdeveloped for energy materials and their integration into operating devices. Specific challenges that have been identified include the lack of uniformity among ontologies and standards; poor accessibility of key datasets for AI/ML based training and testing purposes (*e.g.*, electrochemical testing data, imaging data, physicochemical property data); lack of interchangeability and comparability of data; lack of adaptability to different application areas, and lack of standardised formulation of specimen or cell formats for fully characterised benchmarks.

The DECODE platform will seamlessly chain up modelling and characterisation tools to rationalize the emergence of structure-property relations during fabrication and connect the properties with device level performance and durability. To achieve this, the agnostic DECODE decentralised lab platform will be equipped with 3 integrated innovative methodologies (FABRIC, IRL and CPU) to categorize, down-select, score, map and integrate modelling and characterisation tools and techniques, which will be then demonstrated for a specific use case. Once the benefits of harmonization and standardisation have been established and their consistency with existing data assets is warranted, labs will be more inclined to adopt a new test specimen. DECODE will develop and bring online the tooling for **data standardisation**, including ontologies for data assets from various sources. To warrant security and full control over distribution and sharing of local data, APIs are developed for each node, and in some cases for individual equipment, with full control over tokenization and access by data owners. The data curation, typology of various data sources, standardisation, categorization, ontology and model management, including data access to third parties, will be performed at each partners' discretion, using a Virtual Materials Intelligence platform (ViMi) that is already being developed and maintained by FZJ.

Ontologies and FAIR metadata management in DECODE

There is a lack of standardization of ontologies in energy materials due to diverse domains and applications, heterogeneity in data sources, and lack of universal terminology. Implementing and maintaining ontologies may require sophisticated technological infrastructure and computational

tools. The lack of standardized tools and platforms for ontology development and integration can hinder widespread adoption. DECODE will bring substantial task activities in WP3 to address the lack of standardization in ontologies which will include the development of guidelines by organizations such as ISO, but achieving widespread adoption and agreement within the energy materials, green hydrogen and relevant communities. Finally, updates, revisions, or modifications to existing ontologies will be implemented to accommodate new knowledge or changing requirements. This approach will resolve compatibility issues, especially when integrating data across task and work packages that use different versions of ontologies from materials to devices.[5]

The DECODE platform will use a domain ontology based on EMMC-CSA and adhere to MODA and CHADA principles for classification and documentation of materials modelling and characterisation data. Metadata management will enable users to locate data more quickly and ensure that data quality factors such as consistency, integrity, correctness, and completeness are considered. An operational workflow requires ontology to increase interoperability for the purpose of digital representation. Such a framework will reduce the dimensionality of data, eliminate duplications, and enable data provenance, ease of data generation, dissemination, and redistribution. To enforce FAIRness, ontologies provide semantic context to data and make the interpretations unambiguous[6]. They are already utilised in several projects in materials research[7] and biotech.[8] By using EMMO and its extensions such as CHAMEO, the DECODE ontology will be interoperable with other domain ontologies, such as BattINFO.[9] Several DECODE partners have long expertise in developing data-driven workflows which comply with FAIRness.[10] These are coordinating and participating in DECODE **WP3** and, therefore, ensuring alignment with the objectives and activities carried out there. A few of DECODE's partners will provide their expertise in adopting metadata standards into digital templates and the necessary ontologies to be used in DECODE by the materials experts.

Data ecosystem in DECODE project

This section provides examples of implemented or isolated data-driven models and modules in addressing those challenges related to big data management, energy materials data analytics, AI cloud software infrastructure, and de-centralizing cloud labs for modeling-characterization data pipelines. These approaches are briefly described against challenges in data standardization and *ontology, accessibility* of interchangeable and comparable datasets encompassing the full range of data structures (electrochemistry, performance, component fabrication, and image data for uniformity of fabrication processes)*,* interoperability of data spaces including *workflow integration, non-standardized* specimen, and the variations in innovative MEA cell/stack designs. The DECODE data architecture, Figure 1, warrants the FAIRification of data, well-integrated data flow and management at all levels to support the DECODE community with standard data for accelerating materials-to-device integration.

[5] M. Dreger et al. Materials Informatics, 2023

[6] A. Guizzardi et al., 2021; M. Scheffler et al., 2022

[7] L. M. Ghiringhelli et al., 2021; S. Clark et al., 2022

[8] Gene Ontology Consortium, 2004

[9] S. Clark et al., 2022

[10] Dreger et al., 2023; ViMiLabs; LAGO Thematic service.

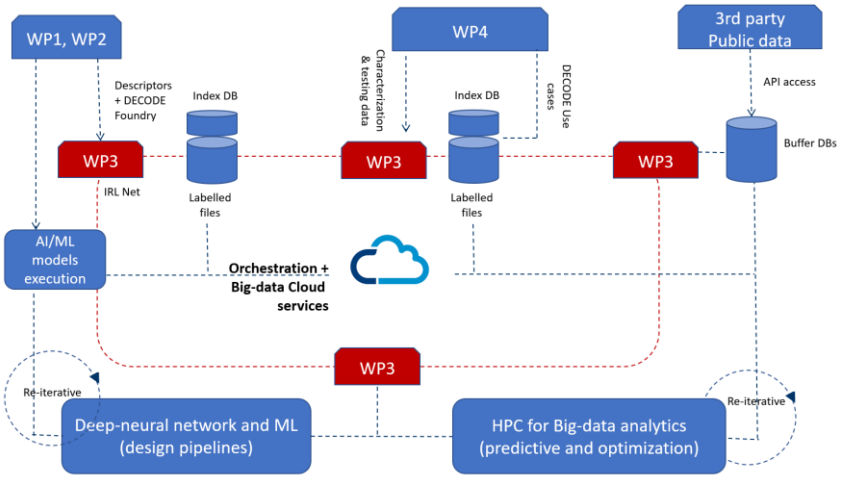

Figure 1. Schematic relation between DECODE modules, database architecture and work packages.

Data management architecture

Despite considerable recent interests to clean energy technologies, adoption of new materials and components for energy storage remains relatively slow. For the most part, this is attributed to the fact that integration of a new or modified material or component into the working environment of a functional device and assessment of its performance is a tedious and time-consuming process. This process clearly lies beyond the capabilities of academic labs, where new materials often originate, and it is not within the purview of the industry. For large-scale industrial uptake of cost-effective materials, there is thus a lack of a seamlessly connected and integrated approach to fully realize the value of data and knowledge in modelling, simulation, and lab-scale characterization techniques. The lack of standard methods for data collection, physical and model-driven analytics, and testing or characterization procedures results in low utilization of data assets from scale-to-scale and lab-to-lab. This shortcoming subsequently leads to inflating development times, duplications, and ultimately increases costs of new materials integration. The lack of standard meta-data and ontologies – the blueprint of connecting field-specific datasets- does not merely limit to technical laboratory data. Indeed, the meta-data standardization needs to be addressed for the entire H_2 value chain ecosystem across products, enterprises, and technologies.

Collecting technical and personal data

Technical and personal data is collected to guarantee the success and high quality of the deliverables in DECODE. The type of the data collected will be text and numeric and overall machine readable, while the format will be mainly plain text (.txt), .csv, images, .xlsx, PDF, .doc, .md, Jason and others in decrypted or encrypted formats. The re-use of existing data is planned in DECODE. The data will be collected from various sources, including work packages, external publicly available or third-party data with written agreement, data generated within WPs and the existing partners' data. The relationship between these data sources, work packages and cloud-enabled platform is schematically shown in Figure 1. The expected size of technical or scientific data is for single files in the range of some KB up to few hundred GBs to TBs. The overall size of personal data is not expected to exceed the MB range.

The data will be useful for the DECODE consortium members, and extended community of energy materials and hydrogen technologies.

Scientific data in DECODE

Scientific data is generated/collected in the frame of DECODE platform and its modules (Foundry, IRL-FABRIC-net, orchestrator) for the development of new materials for devices and their applications in hydrogen usage (fuel cells) or productions (electrolysers). This data includes data generated/collected during the research in DECODE project or generated/collected and published previously by DECODE partners.

In preparation for this initial DMP, an interactive questionnaire addressing all relevant methods and tools was distributed, where partners gathered and uploaded their already published methods and tools in a secure data repository. A knowledge graph was extracted from such data as the first attempt towards building DECODE FAIR meta-data and initiate architecture of the DECODE Foundry. The resulting master knowledge graph will be an integral part of the DMP. From such knowledge graph one can obtain a complete record of all data sets generated/collected from the methods and tools in DECODE. It also ensures that all data is FAIR and follows same meta-data structure.

Computer codes, algorithms, flowcharts

Computer codes, algorithms and flowcharts are generated/collected in the frame of DECODE work packages and DECODE modules to ensure that original codes are available for reproducibility, transferability, and improvement of the existing algorithms. The uniform and finable repositories for codes, flowcharts and alike such as GitHub or gitlab provide a uniform and finable medium for accessibility to original codes and analysis methods that are used by various partners or tasks. This data is generated/collected within WP3. The type and format of the data generated/collected will be either in a machine-readable format or computer codes such as Python, PHP, or similar. The complete documentation will be included to explain the step-by-step flow charts and algorithms. Computer codes and algorithms can be re-used, integrated, modified or improved further. The codes can be the extension of derivatives of the existing codes or a stand-alone code or algorithm. The data will be useful to all partners for analysis or complementing experimental data.

Process know-how

Process know-how is generated/collected in the frame of DECODE work packages and DECODE modules to ensure original codes are available for reproducibility, transferability, and improvement of the existing algorithms. It provides a uniform and finable medium for accessibility to original codes and analysis methods that are used by various partners or tasks also related to new methods for materials characterisation, component testing, component or cell lifetime testing, high throughput materials experimentation or self-driving labs operation for new materials and fuel cell or electrolyser test cells. systems. This data is generated/collected within all technical work packages (WP1, WP2, WP3 and WP4). The type and format of the data generated/collected will be process flow or diagrams and the origin is algorithms or known experimental design or modelling workflow. The data will be useful to testing and characterization labs, materials integrations or technology exploitation.

Scientific publications/presentations/patents

Scientific publications/presentations and patents for dissemination and exploitation of scientific results will be generated to increase the visibility and support the impact of the DECODE project and its exploitable results. These documents will be generated mainly within WP5 by the coordinator, but

also within WP3 and WP4 and disseminated and communicated with the support of WP6. The type of the data collected will be text, numeric and images PDF, .pptx, .xlsx, XPS, and other relevant formats. Existing data may be re-used in WP3 in form of historical datasets for the purpose of training algorithms and for further data enrichment. Scientific publications/presentations/patents will be generated by scientists. The expected size is from few KBs to multi-TBs.

Modelling data and developed algorithms or methodology

This categories of collected data are designed in a way to understand and assess the technical properties of the materials and integration readiness level of materials to devices, device function and behaviour of newly integrated materials in an application / use case context (WP4). WP3 and WP4 are highly aligned with the objective of the project where they are specifically focused on accelerating development of new materials integration with improved properties for hydrogen technologies at the device level. In alignment with the project objectives part of this data is meant to be shared, provided also as a service to other relevant EU funded projects. Materials characterization, component fabrication, component testing, cell testing or modelling data within published or un-published repositories as well as partners' competences and foreground will be used. The expected size ranges from MB (basic models) to TB (large scale assessments; material simulation). The data will be useful for universities, research institutes, materials developers, component developers, lab-scale fuel cell or electrolyser cell testing, device testing and cell / component manufacturers and Small and Medium-Sized Enterprises (SMEs, technology development, automation solutions, consultants, component manufacturers).

Data for sustainability assessment

This data comprises open access primary data and data from cooperative projects collected within WP3 and WP4 to supply a starting point and gather a comprehensive overview in the field of sustainability assessment of materials to device integrations as well as guidelines, key impacts and performance indicators and business proposition canvas templates generated within DECODE exploitation report. All data used will have clear license information in relevant documents. The type of the data collected will be text and numeric and the format will be zolca, xls, ecospold 1 and 2 ILCD for LCA, EXIOBASE for MRIO assessment and xls for LCC and .pptx or .pdf for business intelligence and assessments.

Data management methodology in DECODE

Traditional data management often entails research teams crafting bespoke data infrastructure for specific projects, leading to isolated datasets across different groups. However, for effective machine learning (ML) applications, a larger, more integrated dataset is crucial. The creation of such unified datasets requires merging of these disparate datasets, a task that poses significant challenges owing to the lack of standardization. This is where FAIR principles and ontologies have become indispensable. FAIR principles advocating for data to be findable, accessible, interoperable, and reusable provide guidelines for data management that can bridge the divide between isolated datasets. They enhance scientific reproducibility and accessibility to curated data and facilitate the use of data-driven methods while promoting standardization, which is crucial for data integration.

Role of ontologies

As formal knowledge representations, they are key to making varied metadata machine readable. Using classes, properties, relationships, and rules, they standardized knowledge representation within specific fields, facilitating data and knowledge exchange among researchers. In the interdisciplinary

realm of materials science, standardization through ontologies is crucial. A notable example is the European Materials Modelling Ontology (EMMO)[11], which is a comprehensive ontology developed by the European Materials Modelling Council. EMMO's structure comprises three levels to define real-world objects, specific perspectives, and domain-specific ontologies. It uniquely represents the fabrication, characterization, and simulation of materials and devices. Other projects, such as NOMAD[12], CHAMEO[13], and BigMap[14], use EMMO-based ontologies, enhance their creation process, and improve interoperability owing to shared structures and rule sets. Similar to languages, the power of an ontology is tied to its adoption level, and EMMO is providing a widely accepted foundation for many materials science domains.

FAIR-compliant database

The FAIRness of technical data lies at the heart of data infrastructure, storing fabrication, simulation, and measurement data alongside comprehensive metadata for reproducibility. Currently, the industry standard is relational databases such as MySQL or Postgres, which use tables to represent real-world objects and their instances. However, their structured approach can be rigid, and processing relationships using foreign keys can be slow. NoSQL databases offer a more dynamic data model, including document-oriented, key-value, and graph databases. Document-oriented databases store data in various encodings, allowing for a more flexible model, but they still experience inefficient handling of relationships. Key-value databases offer flexibility but may require extensions for complex use cases. In this study, we used Neo4j, which is a native graph database that uses graph theory for data representation and storage. These databases efficiently handle complex domains, eliminating layers of abstraction between the real world and data model. The nodes and relationships in Neo4j can have properties, and directed relationships offer more context. Graph databases are particularly beneficial in materials research due to the heterogeneity of data. For example, fabrication workflows can be represented naturally by graphs. Unlike table-based processes, graph databases store workflows as node sequences, making variations easy to accommodate without altering existing data. This flexibility allows for nonprescriptive process structuring, which is crucial in the constantly evolving field of materials science.

Data models in DECODE

Our proposed data model utilizes node sequences with labels corresponding to Python classes defined in the neomodel and within the ontology, providing a structure within the database. Nodes can carry multiple labels that signify specific entities with attributes held as key-value pairs to identify them. Label indexing can accelerate node retrieval, but it may reduce write efficiency and increase storage requirements. In materials science, the attribute space for materials, components, devices, and processes is vast and increasing. To manage this, physical quantities are stored in external properties and parameter nodes, facilitating the efficient querying of similar properties or fabrication parameters. This externalization allows the storage of data as scalar or scalar arrays. Neo4j's data structure enables the efficient traversal of relationships with O(1) complexity. Adjacent nodes are physically stored close together to increase the query efficiency, with the database architecture optimized for frequent queries. Additionally, caching during queries enhances the efficiency of the reading, writing, and matching operations.

[11] https://emmc.eu/wp-content/uploads/2021/06/Euronano2019_MODA_CHADA-min.pdf

[12] https://nomad-lab.eu/nomad-lab/

[13] https://emmc.eu/wp-content/uploads/2021/06/Euronano2019_MODA_CHADA-min.pdf

[14] https://www.big-map.eu/

We have developed a novel data model that serves as the foundation for dynamic data infrastructure tailored to the fabrication, measurement, and simulation of energy materials. This data model accommodates workflows and processes of any complexity, and can be adapted to include new materials, components, or processes. To represent the domain of hydrogen technology, particularly fabrication and characterization, we introduced an ontology based on the EMMO. Data are stored in the native graph database neo4j, which boasts efficient traversal of the fabrication processes. The distinctiveness of our model lies in its flexibility in depicting the fabrication workflows and measurements. Its strongest application is in automated laboratories, which requires efficient data management. Current automated laboratories often construct unique data management systems, resulting in isolated, non-standardized data repositories. Our model offers a solution to this fragmentation problem, as it can represent any workflow in a great detail.

The adoption of neo4j as a native graph database aligns with the data structure that we aim to capture. For progressive data-driven workflow optimization, the integration of data FAIRification in experimental workflows is crucial. Notably, the fabrication process data often lack FAIR characteristics owing to its heterogeneous nature and the absence of standardization workflows. Transitioning from the traditional, rigid, table-based data model to our flexible, graph-based model facilitates the storage of FAIR- and AI-compatible data. This access to FAIR experimental data could catalyze the adoption of data-based techniques such as transfer learning [3], Figure 1.

Models Deployment

Model dockers need to be built on certain standards with proper documentation and associated architectures. Metadata standards can be implemented using graph database management module that matches datasets to specific ontologies [2,3]. The seamless integration of model management with other orchestration units will ensure that correct information is sent to the data control unit to schedule or execute proper models for data analytics, inverse-design or further device integration.

Databases and meta-data standards

The metadata management system will be utilized in the existing ViMi Labs platforms at FZJ. ViMi Labs is being developed at FZJ which has a peer-to-peer (P2P) decentralized architecture with cross-dimensional operation capabilities, data manipulation tools for inductive and deductive analysis, and multiuser support; its functionalities have been demonstrated in multi-source and heterogeneous data acquisition for the development of electrocatalyst materials. The client/server architecture of ViMi allows the implementation of metadata management system.

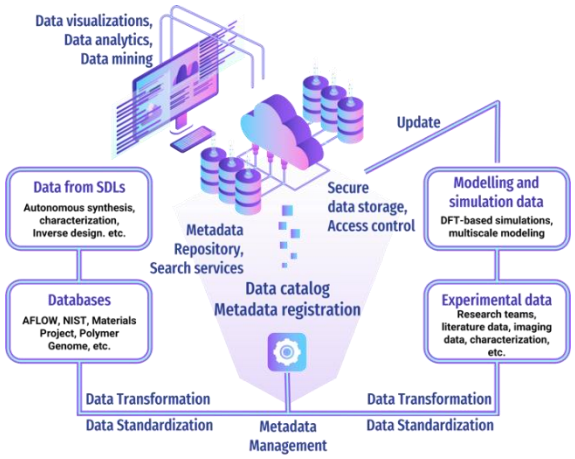

Figure 2. Overview of the general architecture of the analytical data store system in this project

In the approach proposed in this DM report, we underscore the necessity for standardized and meticulously documented data from testing procedures. This approach was designed to yield high-quality benchmark data, which form the bedrock for model validation. We propose the incorporation of a broad array of characterization methodologies and tools. These include various tomography techniques characterized by distinct spatial resolutions and fields of view such as X-ray CT, FIB/SEM, and electron tomography. In situ and operando spectroscopic and diffraction methods such as XANES, EXAFS, SAXS, XPS, XRD, DEMS, and FTIR data will be utilized to delve into the binding energies, oxidation states, and atomic-scale coordination of the catalyst materials.

Furthermore, imaging methods, including neutron radiography, TEM, and XTM, will be employed to illuminate the two-phase liquid-gas saturation in the porous media of functioning cells and the ensuing impact on transport processes. Traditional electrochemical methods, such as polarization curves, cyclic voltammetry, and electrochemical impedance spectroscopy, are also part of the methodological framework for evaluating the performance. We propose the use of a comprehensive ontology for the multiscale characterization of electrochemical devices, informed by the principles of EMMO and CHADA. This ontology is envisaged to span from nanoparticles through complex interfaces and nanopores to ink agglomerates, all the way up to the cellular level.

Within ViMi Labs platform [2], we have established a collaborative imaging platform that in future uses semantic search capabilities. This platform facilitates automated image labelling, standardized metadata record collection, image quality control, pre-processing, and the generation of unique data identifiers for traceability. The annotated imaging datasets generated through this process will be utilized to construct data-driven models for autonomous characterization by employing end-to-end deep-learning-based pipelines. These pipelines harness deep learning techniques based on Convolutional Neural Networks to perform a variety of analytical tasks, such as segmentation, detection, and super-resolution microscopy.

Ethical aspects and personal data in DECODE

Ethical issues related to the DMP are discussed in D5.1 (Ethics requirements). This section (directly copied from D5.1) clarifies what constitutes personal data in DECODE project and the different types that may be encountered in this project. In addition, it discusses the ethical principles for managing personal data, including confidentiality and respect for privacy, consent and anonymization and data protection impact assessment.

Types of Personal Data in DECODE project

Personal data refers to any information relating to an identified or identifiable natural person. This can include direct identifiers such as names and addresses, as well as indirect identifiers like IP addresses or pseudonyms that, when combined with other data, can identify an individual. The General Data Protection Regulation (GDPR) of the European Union provides a comprehensive definition of personal data, emphasizing that it encompasses any information that can be used to identify a person directly or indirectly.[15]

Direct Identifiers: These are data points that explicitly identify an individual, such as:

Names
Addresses
Email addresses
Phone numbers
Indirect Identifiers: These data points do not directly identify an individual but can be used in combination with other data to ascertain their identity. Examples include:

IP addresses
Geographical locations
Dates of birth
Pseudonyms
Sensitive Personal Data: Also known as special category data under GDPR, this includes information that requires additional protection due to its sensitive nature, such as:

Racial or ethnic origin
Political opinions
Religious or philosophical beliefs
Genetic data
Biometric data
Health information
Sexual orientation

Behavioral Data: This includes data derived from an individual's behavior, which can be collected through various means such as:

Online browsing habits
Purchase history
Social media activity

Understanding these different types of personal data is crucial for DECODE researchers to manage and protect the data appropriately throughout the research lifecycle.

Ethical Principles for Managing Personal Data

Ethical handling of personal data in DECODE project involves adhering to principles that protect

[15] European data protection regulations (GDPR), General Legal Text, https://gdpr-info.eu/

individuals' rights and maintain trust in the research process. Key principles include:

Confidentiality: Ensuring that personal data is accessible only to authorized individuals and entities. Confidentiality protects the privacy of research participants and maintains their trust. Researchers should implement strict access controls and use secure data storage solutions to maintain confidentiality.

Respect for Privacy: Researchers must respect participants' privacy by collecting only the data necessary for the research and using it in ways that align with participants' consent. This includes avoiding unnecessary intrusion into participants' private lives and ensuring that data collection methods are minimally invasive.

Transparency: Being transparent with participants about how their data will be used, stored, and shared. This involves providing clear, comprehensive information at the time of data collection and maintaining open communication throughout the research process.

Data Minimization: Collecting only the minimum amount of data necessary to achieve the research objectives. This principle helps reduce the risk of privacy breaches and ensures that the data collection process is proportionate to the research needs.

Fairness and Accountability: Ensuring that data processing is FAIR (Findable, Accessible, Inter-operable and Reproducible), lawful, and accountable. Researchers must comply with relevant legal and ethical standards and be prepared to demonstrate their adherence to these standards through documentation and regular audits.

Consent and Anonymization

Obtaining informed consent is a fundamental ethical requirement for collecting personal data in DECODE project. Informed consent involves providing participants with clear, comprehensive information about the research and obtaining their voluntary agreement to participate. Key elements of informed consent include:

Information Provision: Participants will be fully informed about the purpose of the research, the types of data to be collected, how the data will be used, and any potential risks or benefits. Information should be provided in a clear and accessible manner.

Voluntariness: Participation in the consent will be voluntary, with participants able to withdraw their consent at any time without penalty. Researchers should ensure that participants do not feel coerced or pressured to participate.

Documentation: Consent will be documented, either through written consent forms or, in some cases, electronically. This documentation serves as evidence that participants have agreed to the terms of the research.

Anonymization involves transforming personal data so that individuals cannot be identified, either directly or indirectly. This is a crucial step in protecting participants' privacy, particularly when sharing data with third parties or making it publicly available. Common methods of anonymization include:

Pseudonymization: Replacing direct identifiers with pseudonyms or codes. While pseudonymized data can be re-identified with additional information, it reduces the risk of immediate identification.

Aggregation: Combining data from multiple individuals to produce summary statistics or group-level information. This reduces the likelihood that individual data points can be linked to specific individuals.

Data Masking: Altering data to obscure identifying characteristics, such as through generalization or randomization (e.g., introducing random noise to data points).

Suppression: Removing certain data points or identifiers entirely, particularly when they are not essential to the research analysis.

By implementing these anonymization methods, researchers can significantly reduce the risk of privacy breaches while still enabling valuable data analysis and sharing.

Data Protection Impact Assessment (DPIA)

A Data Protection Impact Assessment (DPIA) is a process designed to identify and mitigate risks associated with processing personal data. It is particularly important for projects that involve large-scale data processing, sensitive technical or personal data, or high-risk data processing activities.

We will determine whether a DPIA is required based on the nature and scope of the data processing activities in DECODE project. GDPR mandates a DPIA for processing activities that are likely to result in a high risk to individuals' rights and freedoms. As such, we do not anticipate DPIA is required in the case of DECODE project.

Nevertheless, DECODE will Provide a detailed description of the data processing activities, including the types of data to be collected, the purposes of processing, and the methods used for data collection, storage, and sharing. Moreover, we will evaluate whether the data processing is necessary and proportionate to achieve the research objectives. Potential risks will be identified to data subjects' rights and freedoms, such as risks to privacy, data security, and potential misuse of data. The implement measures will help mitigate identified risks. This can include technical measures (e.g., encryption, access controls) and organizational measures (e.g., training, policies, and procedures). More importantly, we will engage with DECODE stakeholders, including data protection officers with different partners, ethics committees, advisory board members, and, where appropriate, research participants, to gather feedback and ensure that all relevant perspectives are considered.

We plan to regularly review and update the DPIA to reflect changes in data processing activities or the emergence of new risks. Moreover, continuous monitoring and review help ensure ongoing compliance with data protection standards.

In conclusion, managing personal and secure technical data in research requires a comprehensive understanding of what constitutes personal data, adherence to ethical principles, obtaining informed consent, employing anonymization techniques, and conducting thorough DPIAs. These practices ensure that personal data is handled responsibly, protecting the rights and privacy of research participants while facilitating valuable scientific inquiry.

Open access to scientific publications

According to the Article 4.5 in CA, each beneficiary must ensure open access (free of charge online access for any user) to all peer-reviewed scientific publications relating to its results. Where necessary, the Parties shall cooperate in order to enable one another to fulfil legal obligations arising under

applicable data protection laws (the Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data and relevant national data protection law applicable to said Party) within the scope of the performance and administration of the Project and of this Consortium Agreement. In particular, the Parties shall, where necessary, conclude a separate data processing, data sharing and/or joint controller agreement before any data processing or data sharing takes place.

Open access to research data

Regarding the digital research data generated in the action ('data'), the beneficiaries will deposit in a research data repository and take measures to make it possible for third parties to access, mine, exploit, reproduce and disseminate — free of charge for any user — the following:

- the data, including associated metadata, needed to validate the results presented in scientific publications, as soon as possible
- other data, including associated metadata, as specified and within the deadlines laid down in the DMP
- provide information — via the repository — about tools and instruments at the disposal of the beneficiaries and necessary for validating the results (and — where possible — provide the tools and instruments themselves).

This does not change the obligation to protect results according to CA, the confidentiality obligations, the security obligations or the obligations to protect personal data, all of which still apply in accordance with CA. In case any information provided conflicts with the GA, CA or with the Commission, the terms of the latter shall prevail.

Data security

Security is an essential aspect in DECODE project. Establishing secure channels for communication between users and processes including authentication and data confidentiality, as well as access control and authorization, require design protocols to enforce certain security policies. DECODE will implement state-of the art organizational and technical security measures to ensure compliance with data protection regulations and to prevent unauthorised access to personal data or to the equipment used for processing. This is carried out at all level, from module development to AI model training to module deployment in ViMi Labs platform and user dashboard access. To ensure data protection during data transmission, state-of-the-art encryption technology is used. Security mechanisms will be applied to manage the distribution of cryptographic keys as well as for group authorization in order to mitigate risks such as cyberattacks, data loss, service interruption, and malware programs.

Key and authorization management. Public-key certificates will be used for the distribution of cryptographic keys. The channel to send the public key provides authentication, while the paired private key is sent from a secure channel with both authentication and confidentiality (Figure 3). Additionally, the secure access delegation will be done through Open Authorization (OAuth 2.0) protocol using tokens of user defined expiry length.

Cloud-compatibility and scalability. This task will focus on the inter-operability of the cloud communication protocols, including data security, metadata definitions, and scalability of the FAIR meta-data framework. FZJ will lead the development of the secured cloud communication and scalability tests. Moreover, FZJ will lead the development of dynamic data management approaches and integration with ontologies and data infrastructure. ViMi Labs will assist in the process of down-selection and assessment of the security protocols specific to the use cases.

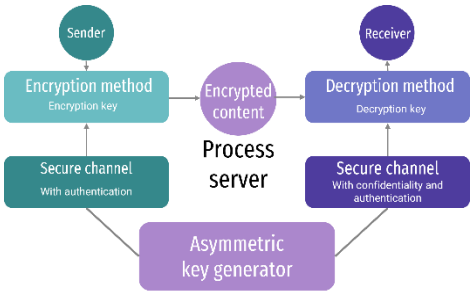

Figure 3. Mechanism for public key distribution

Creating AI-ready scientific data in DECODE

Scientific research at its core, is a data-driven process where heterogeneous sets of information, i.e. research data (documents, text, spreadsheets, lab notebooks, images, models, algorithms, scripts, methods, know-how, workflows, operating procedures and protocols, and software (containers)) are utilized for creating or exchanging ideas, interpretation of the results, and examination of the hypotheses. Scientific data is also a fundamental element for building and training strong and reliable AI/ML models. The quality, applicability, and trustworthiness of the pre-trained AI models in DECODE project and recyclability of the AI-generated data for re-training purposes are decisively relying on accuracy and reliability of the employed training methodologies. Yet, resulting information and thus accuracy of the knowledge can be greatly impacted by inherent interdependencies among underlying data and training pipelines. This makes AI-ready data management in DECODE an important requisite for creating state-of-the-art methods for creating DECODE modules and ultimately the DECODE platform.

In AI/ML model implementation, data is at the beginning of the training pipeline, processing specific requirements to maximize their utilities in view of accessibility, meta-data availability, reproducibility, and inter-operability. In any case the preparation of well-labelled test data and underlying practical methods and reliable tools for research data management are highly crucial during the AI/ML training. This relates to aspects such as the quality and quantity of the raw data (e.g., accurate, relevant, diverse, and error-free), the quality and quantity of data labels with relevant metadata (e.g., consistent, community-accepted, interoperable annotations), technical considerations (e.g., accessibility and interoperability of (meta)data sources, scalable/efficient storage, or compatibility of data formats with chosen AI/ML methods). In addition, the processes for generating or labelling training data need to allow for reproducibility and auditability, Figure 4.

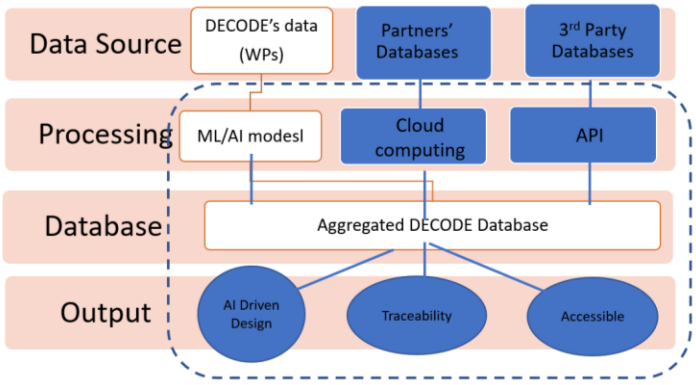

Figure 4. The processes for generating or labelling training data in DECODE

The following activities are particularly considered to ensure high quality AI-ready data in DECODE project:

Cloud- enabled models deployment platforms: The standardised model dockers, user friendly, and enterprise-grade platform are built using ViMi Labs platform to encourage usage and module adoption. This needs to be built on certain standards with proper documentation and associated architectures. Metadata standards can be implemented using graph database management module that matches datasets to specific ontologies. The seamless integration of model management with other orchestration units will ensure that correct information is sent to the data control unit to schedule or execute proper models for data analytics, inverse-design, or further device integration.

Automated data ingestion: Technical table data can be ingested in a semi-automated fashion by a pipeline and agnostic to the table structure, terminology, or size. To start the transformation process, the table is parsed, and a dictionary that contains the headings, a few sample rows, and a string that provides additional context is generated and fed to the pipeline. DECODE will adopt modern and intelligent data management leveraging ontologies and novel data models with LLM-enabled semi-automate data mapping and data ingestion

Implementing semantic search: Current data resources share a common logic. The use of Semantic Web technologies and fine-tuned LLMs **ensures** uniform labelling of data assets as crucial parts of a semantic search pipeline. LLMs are used to generate descriptions and alternative labels for ontology classes. These labels and descriptions are used to generate embeddings (vector representations of the human language). The generated embeddings are linked to the ontology classes within the database.

Creating ontologies pool: DECODE ensures standardization through adoption of ontologies. As formal knowledge representations, they are key to making varied metadata machine readable. Using classes, properties, relationships, and rules, they standardized knowledge representation within specific fields, facilitating data and knowledge exchange among DECODE partners.